Name: Achyut Shukla Batch: ML – 1 PRN: 20070122005

॥वसुधैव कुटुम्बकम्॥

**SYMBIOSIS INTERNATIONAL (DEEMED UNIVERSITY)**

(Established under section 3 of the UGC Act 1956)

**Re - accredited by NAAC with 'A' Grade**

**Founder: Prof.Dr. S. B. Mujumdar, M.Sc.,Ph.D. (Awarded Padma Bhushan and Padma Shri by President of India)**

---

# Lab Assignment – 3

**Aim :**

Implement Decision Tree Algorithm on a sample case study and data set. Evaluate Results.

**PART – A**

Theory of Decision Trees

A decision tree is a powerful and widely used machine learning algorithm that is used for both classification and regression tasks. It's a supervised learning method that visually represents decisions and their possible consequences in a tree-like structure. Each internal node of the tree represents a decision based on a particular feature or attribute, and each leaf node represents a class label or a predicted value.

Components of a Decision Tree:

1. Root Node: The topmost node in a decision tree, representing the initial decision or attribute that best separates the data.

2. Internal Nodes: Nodes that represent decisions based on specific features or attributes. Each internal node splits the data into subsets based on a particular feature's value.

3. Leaf Nodes: Terminal nodes that indicate the final outcome or prediction. In classification, each leaf node corresponds to a class label, while in regression, it holds a predicted continuous value.

4. Branches: Edges connecting nodes, indicating the flow of decisions based on attribute values.

Working Principle:

The construction of a decision tree involves recursive partitioning of the dataset into subsets. The goal is to create segments that are as pure as possible in terms of the target variable. For classification tasks, purity means that the majority of data points in a subset belong to the same

class. For regression tasks, purity implies minimizing the variance of the predicted values within the subset.

The algorithm uses various techniques to determine which attribute to split on at each node. Some common methods include:

- Gini Impurity: Measures the probability of a randomly selected element being misclassified. It favors attributes that result in a more uniform distribution of classes.

- Entropy: Measures the impurity or disorder of a set of examples. It seeks to maximize the information gain at each split.

- Information Gain: The reduction in entropy or impurity achieved by a particular split. It selects the attribute that provides the most significant information gain.

- Gain Ratio: Adjusts information gain by considering the intrinsic information of each attribute, reducing the bias towards attributes with many values.

- Chi-Square: Determines if the distribution of class labels is significantly different in different subsets.

Advantages of Decision Trees:

1. Interpretability: Decision trees provide human-readable, intuitive representations of decisions and their outcomes, making them easy to understand and visualize.

2. Handling Non-linearity: Decision trees can effectively handle non-linear relationships between features and target variables.

3. Feature Importance: Decision trees allow for the assessment of feature importance, helping to identify the most influential attributes in making decisions.

4. Handling Missing Values: Decision trees can handle missing values by creating surrogate splits.

Limitations:

1. Overfitting: Decision trees are prone to overfitting, especially when they become too deep and complex. Pruning techniques and setting depth limits can help mitigate this.

2. Instability: Small variations in the data can result in significantly different decision trees. Techniques like ensemble methods (Random Forests, Gradient Boosting) can help stabilize predictions.

3. Bias towards Dominant Classes: Unbalanced datasets can lead to biased decisions in favor of the dominant class.

4. Discretization: Decision trees work better with discrete data; continuous attributes need to be discretized.

In conclusion, decision trees are versatile algorithms with a clear, interpretable structure. While they have certain limitations, they can be enhanced and combined with other methods to create robust and accurate machine learning models.

**PART – B**

Experiment:

**Step 1** – Study of Dataset (Dataset references are given below)

**Step 2** − Data Pre-processing step

**Step 3** − Fitting a Decision-Tree algorithm to the Training set

**Step 4** – Predicting the test result.

**Step 5** – Test accuracy of the result(Creation of Confusion matrix)

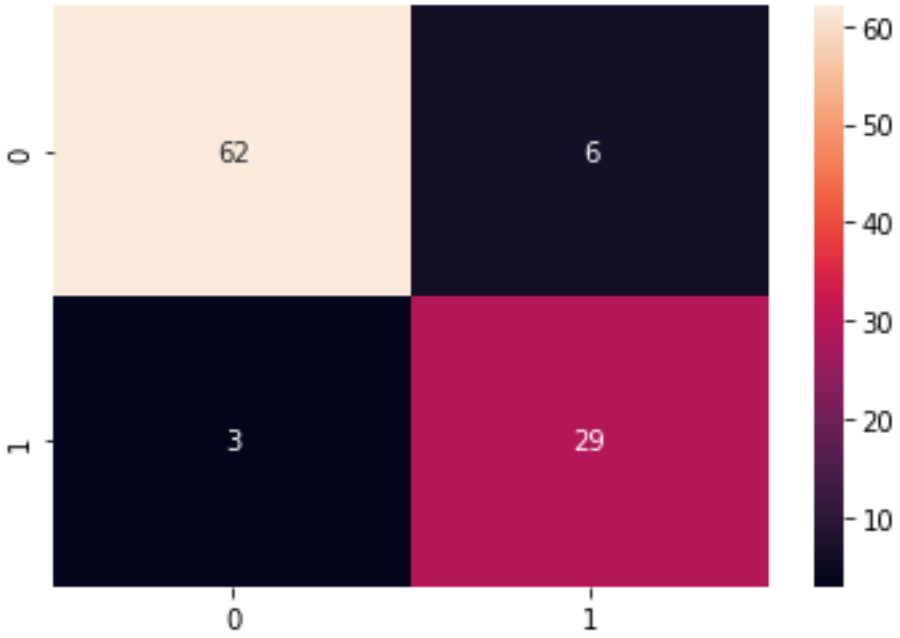
**Step 6** – Visualizing the test set result.

Datasets :

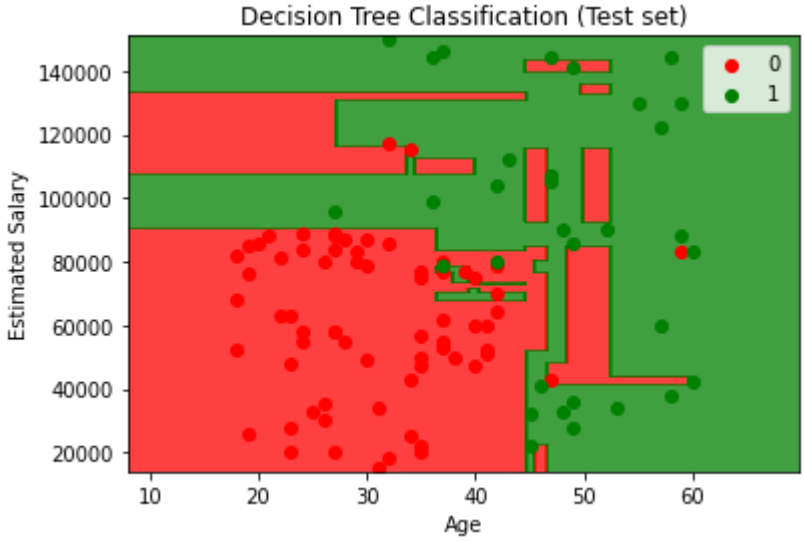
1. https://www.kaggle.com/datasets/rakeshrau/social-network-ads

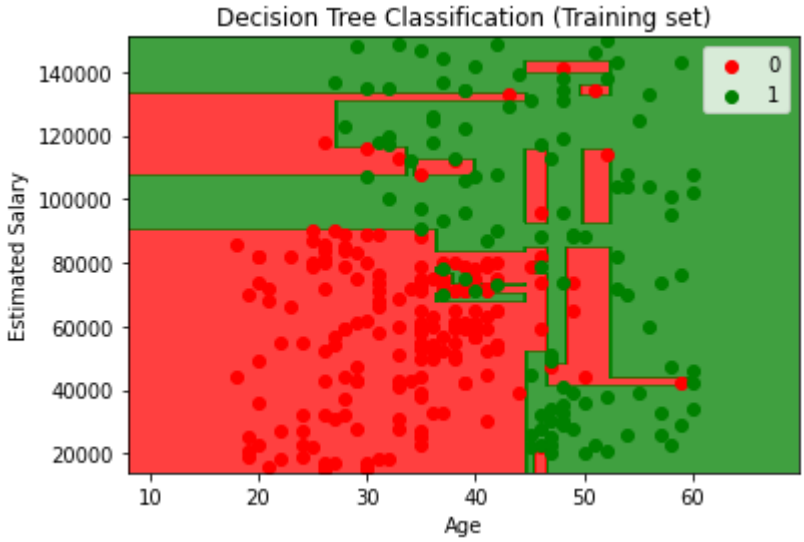
Output:

1. Confusion Matrix




2. Data Visualization




Inference Discussion

Implemented and Understood Decision Tree Classifier algorithm and visualized the data and the output with an accuracy of 100%.